SW

The
PSYCHOLOGICAL
RECORD . . . .

OCTOBER, 1938
Vol. II No. 14

HOMOGENEITY AND HETEROGENEITY IN LANGUAGE; IN ANSWER TO EDWARD L. THORNDIKE

George K. Zipf

THE PRINCIPIA PRESS, INC.
BLOOMINGTON, INDIANA

Price of this number, 50 cents

The Principia Press, Inc., has undertaken the publication of this cooperative journal to afford authors the opportunity of immediate publication at the least possible expense. The present low cost of publication, and possible future reductions, depend entirely upon the number of subscriptions. The subscription price has been made low to induce individuals to subscribe. Under the Articles of Incorporation of the Principia Press no profit can be made on any of its publications. Therefore an increase in the number of subscribers will be reflected in reduced costs to authors and in increase in the number of pages published annually.

EDITORIAL BOARD. The above named board of associate editors have agreed to serve in an active capacity for a period of three years, and all manuscripts will be submitted to them.

MANUSCRIPTS may be sent to the appropriate associate editor or to Dr. J. R. Kantor. Longer papers (30 or more typewritten pages) in the above mentioned fields are especially desired.

COSTS. The cost of publication will be $2.00 per page of approximately 470 words. The charge for line cut illustrations will be $2.00 regardless of size. Special charges will have to be made for complex tabular matter and for half-tone or colored illustrations.

REPRINTS. One hundred copies of each paper will be supplied gratis. Additional copies may be purchased in any quantity at a rate which will be set when the author is informed of the cost of publication.

ROYALTIES. Fifty per cent of the net income from the sale of individual copies, or from copies sold as part of back volumes, will be credited as royalties to the author's account. Royalties cannot be paid on income from subscriptions current in the year of publication.

SUBSCRIPTIONS. The subscription price is $4.00 per year for an annual volume of approximately 500 pages. Individual papers will be sold at an advertised price, which will depend upon the size of the article. Foreign subscriptions will be $4.50.

CORRESPONDENCE concerning editorial matters should be addressed to Dr. J. R. Kantor. Indiana University. Business correspondence should be addressed to Dr. C. M. Louttit, Principia Press, Inc., Bloomington, Ind.

HOMOGENEITY AND HETEROGENEITY IN LANGUAGE; IN ANSWER TO EDWARD L. THORNDIKE*

By George K. Zipf,
Harvard University

In the words of Edward L. Thorndike ('37, p. 399): "Zipf ('35, p. 39f and '37*, p. 239f) has suggested that the well-known but unmeasured facts that there are very few very frequently used words, and increasingly larger numbers of less and less frequently used words, at least within certain limits, can be put in order to a first approximation by the equation ab^2 = a constant for the given field of speech or writing, in which b = any number of occurrences and a = the number of words occurring that number of times." Continuing, Dr. Thorndike ('37, p. 400) reports his own independent investigation: "I have recently completed a count of about 4½ million running words from 120 books taken with few exceptions from those recommended by Terman and Lima for supplementary reading by children in grades 3 to 8." Dr. Thorndike then explains his method of classification and I agree with him that for all essential purposes: "This count satisfies Dr. Zipf's criteria" (ibid.)

Upon inspection of the facts of his count, Dr. Thorndike arrives at conclusions, on three of which I shall now take a stand: (1) "The relation between a and b for all the speech or writing of any person or group of persons may then vary greatly according to the size of the sample that is used. If ab^2 is a constant for words occurring from 1 to 1,000 times in samples of the size used by Zipf for the persons and groups used by Zipf, it almost certainly will not be that for the total usage of those persons and groups" ('37, p. 402f.) The size used by Zipf was of the order of magnitude of about 50,000. (2) "It seems to me unlikely that the relation for such total usage, no matter how uniform it turned out to be, would be evidence of any uniform and ubiquitous tendency toward a certain equilibrium between frequency and variety. I should expect that it would be in some measure a statistical artifact, adding those things which lost much of their instructiveness by being combined" ('37, p. 405). (3) "That there is a force wider and deeper than any of these acting to produce equilibrium for equilibrium's sake, I do not believe. Nor does Dr. Zipf, I think, though his words

* Recommended for publication by Dr. J. R. Kantor, September 15, 1938.

may be so interpreted" ('37, p. 406). My ensuing treatment will be such that none of the three points of Dr. Thorndike will suffer from being removed from their context. To make clear the scope of Dr. Thorndike's three points I shall briefly review the statistical background of our common problem.

Zipf ('35, p. 44f) reported, on the basis of an analysis of Eldridge's frequency count of American newspaper English, the following: if one ranks the different words (in fully inflected form) in the decreasing order of the frequency of their occurrence, one finds that the frequency of any word of rank R is 1/10R of the total number of running words in the sample. Thus, of all occurrences, the most frequent word is 1/10, the next 1/20, the third 1/30, the Rth being 1/10R, or, in general terms:

FORMULA ONE: $F_r = \frac{S}{10R}$

where S (a constant) represents the sum of all the running words in the sample analyzed, and F_r the frequency of occurrence of the word of rank R.

This is an empirical formula, and, Eldridge's study not being generally available, I offer empirical evidence from Prof. Hanley's excellent index and statistical study of James Joyce's *Ulysses* (Hanley '37). In *Ulysses* 29,899 different words have a total occurrence (S) of 260,430. In Table 1 are presented in the column at the left arbitrarily selected observed ranks of different words; in the next column is given their theoretical frequency of occurrence i.e. $\frac{260,000}{10R}$; in the next column at the right the observed frequency of the corresponding rank; and in the last column at the right, the product of observed rank by observed frequency (items in column 1 multiplied by corresponding items in column 3). Thus the 10th most frequent word (R = 10) has a theoretical frequency of 2,600, an observed frequency of 2,653, and an observed product of rank by frequency of 26,530.

The closeness of the approximation of the observed frequencies (Column III) to the theoretical frequencies (Column II) is, I think, apparent and needs no comment. The fact that the products of Column IV also approximate a constant need not surprise us, for we have but multiplied both sides of Formula I by *R* and find:

FORMULA TWO: $RF_r = C$

where $C = \frac{S}{10}$

Formula II ("Rank times frequency is constant" or, here-

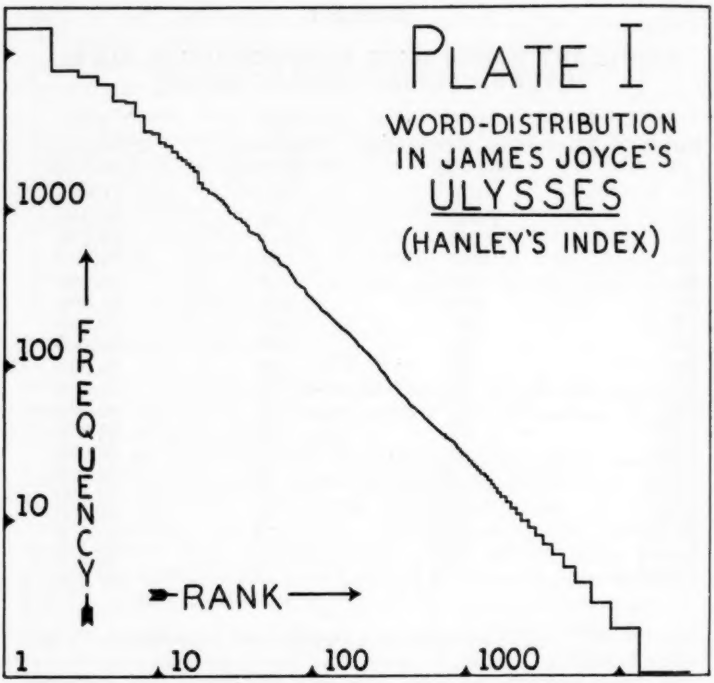
TABLE I

ARBITRARY RANKS WITH FREQUENCIES IN JAMES JOYCE'S *ULYSSES* (HANLEY INDEX)

I	II	III	IV
Rank	Theoretical Frequency	Observed	Product of
(R)	$\frac{260{,}000}{10R}$	Frequency	I and III
		(F)	(FxR)
10	2,600	2,653	26,530
20	1,300	1,311	26,220
30	867	926	27,780
40	650	717	28,680
50	520	556	27,800
100	260	265	26,500
200	130	133	26,600
300	87	84	25,200
400	65	62	24,800
500	52	50	25,000
1,000	26	26	26,000
2,000	13	12	24,000
3,000	8.6	8	24,000
4,000	6.5	6	24,000
5,000	5.2	5	25,000
10,000	2.6	2	20,000
20,000	1.3	1	20,000
29,899	.9	1	29,899

inafter RF = C) describes an equilateral hyperbola. If we alter Formula II to log F, + log R = log C, we may expect our data, when plotted on log — log grid to appear as a straight line of negative slope of one (i.e. to slant descending to the right at an angle of 45°). On Plate I appears the *Ulysses* material thus plotted, with the expected linearity and slope. (We shall ignore in the present paper the slight concavity downwards at the top.) It must, of course, be emphasized that *all vertical lines* of the curve on Plate I are meaningless, being drawn merely for the convenience of the reader's eye. But the horizontal lines (hereinafter *treads*) of the curve at the lower end, to which we now turn, have indeed meaning; they represent different words of like frequency of occurrence. For example, the *tread* at the frequency of 2 represents the ranks of all words occurring twice in *Ulysses* (actually 4776).

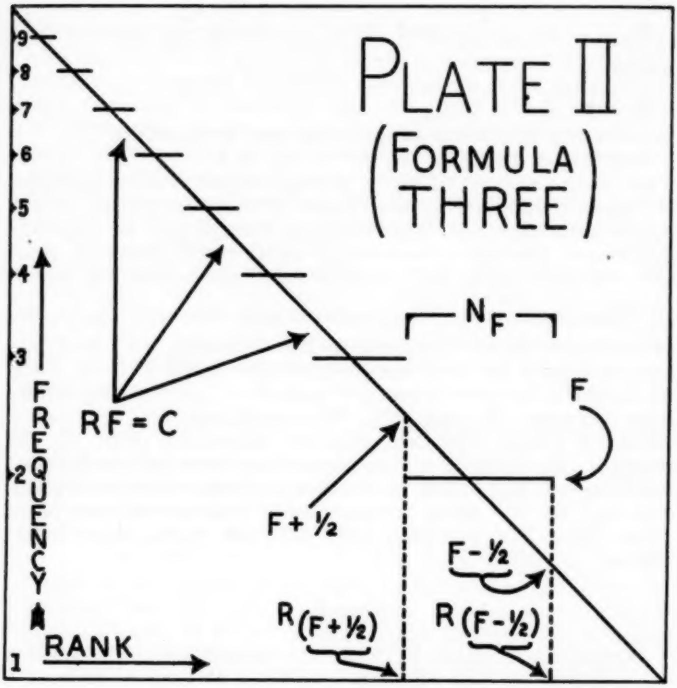
The fact that straight horizontal lines, or *treads*, should appear in the lower frequencies need not amaze us. Formula II, as stated, would call for fractional frequencies of occurrences in any sample (e.g. frequencies of, say, ¼, ½, ¾, 1¼



etc.); since a word cannot actually occur a fractional number of times in any given sample of connected speech, our Formulae I and II really necessitate the appearance of treads of varying size. However, it is reasonable to suppose that our formulae impose restrictions upon the various sizes of the treads, and we shall now see that the approximate size of the treads may be deduced from our formulae.

On Plate II is presented an ideal curve for the lower portion of Formula II, with treads for several of the lowest frequencies; the curve represents the given frequencies for any set of data following Formula II thus plotted. Let us confine our attention to the *tread* marked *F*, an arbitrary frequency; that is, to the number of words (N_F) occurring, say, twice.

Two things are certain: (1) the line RF = C, having a negative slope of 45° will cut all horizontal treads at only



one point; (2) the tread at F = 2 is theoretically a projection on F = 2 of that portion of the line RF = C which falls between F + ½ and F — ½, thus marked. That is, since we may have only an integral number of frequencies, we should expect for each frequency-integral a number of ranks (i.e. number of different words) represented by the ranks on RF = C half way between the integral-frequency above, on the one hand, and the integral-frequency below on the other. Therefore if we subtract the rank of the word at F + ½ from the rank of the word at F — ½, we should have N, the number of words occurring twice, or the size of the tread at F = 2. Thus the number of different words of a given frequency, N_f, can be obtained to a reasonably close approximation by solving the equation RF= C for R at F + ½ and for R at F — ½, and by subtracting the former R from the latter *R*, or

$$R_{(f-\frac{1}{2})} = \frac{C}{F-\frac{1}{2}} \text{ and } R_{(f+\frac{1}{2})} = \frac{C}{F+\frac{1}{2}}$$

hence

$$N_f = \frac{C}{F-\frac{1}{2}} \cdot - \frac{C}{F+\frac{1}{2}}$$

or, clearing fractions, simplifying and transposing:

FORMULA THREE: $N_f\ (F^2 - \frac{1}{4}) = C$

That is, the number of words of like frequency of occurrence, N_f, when multiplied by the square of their frequency minus ¼, remains constant. This formula, though still an approximation, is perhaps a somewhat closer approximation than $ab^2 = C$ (Zipf '35, p. 40f.) mentioned at the beginning, which Dr. Thorndike used. We remember that C is still $\frac{S}{10}$ as explained previously. Incidentally the constant, — ¼, is of importance only for very low frequencies; when F is as much as 10, the difference between F^2 and $F^2 - \frac{1}{4}$ is but the difference between 100 and 99¾. For small samples, this (or a similarly small) constant may be important (Zipf '35, '37 *passim*). In Table II are presented the observed products of $N_f(F^2 - \frac{1}{4})$ for a few of the lower frequencies in *Ulysses,* and also for the lower frequencies of four of Plautus' plays (Zipf '35, p. 27), in which 8,437 different words occur 33,094 times.

TABLE II

FORMULA THREE FITTED TO ARBITRARY LOWER FREQUENCIES IN THE *ULYSSES* AND PLAUTUS MATERIAL

Frequency (F)	Observed $N_f(F^2 - \frac{1}{4})$	
	Ulysses	Plautus
1	12,324	4,075
2	15,410	4,490
3	19,193	4,280
4	20,239	4,750
5	22,424	3,985
6	22,773	4,504
7	23,546	4,241
8	23,651	4,399
9	24,063	4,366
10	22,145	4,289
15	21,576	2,922
20	27,844	5,996
30	18,000	3,600
40	25,600	4,800
50	22,500	5,000

The values for *Ulysses* in Table II increase progressively in size up to F = 9 and then become somewhat variable; we might be tempted to expect a better fit with a small positive constant or function (e.g. $F^2 + F$); and I shall gladly abandon any of the above formulae for ones that fit the data better. Yet the Plautus values are on the whole quite constant, and suggest that our Formula III may stand for the time being at least, particularly since Dr. Thorndike has observed that Formula III (even if corrected with — ¼ according to my inspection of his data) is not valid for his very large sample. Let us therefore *assume* that Formula III is invalid for very large samples, like *Ulysses,* but valid for smaller ones, like that of Plautus, and, in dismissing our data, simply ask why it should apply to magnitudes of 33,000 or of 50,000 and not for magnitudes of 260,000 or more running words.

Now Formula I, our $F_r = \frac{S}{10R}$, represents a harmonic series, and shows where Dr. Thorndike may have made a mistaken interpretation. We note that we can increase indefinitely the general series

$$\frac{1}{10} + \frac{1}{20} + \frac{1}{30} + \ldots + \frac{1}{10R} + \frac{1}{10(R + 1)} \ \ldots$$

simply by ever adding one more to *R*. And as we increase R without limit we also increase the sum of these fractions without limit. If we set a limit to the size of *R* we automatically set a limit to the size of the sum; and conversely, *if we set a limit to the size of the sum, we set a limit to the size of* R. And it happens that, in our formula for language, the size of the sum is limited, as we shall see, and therefore the size of R (i.e. the number of different words in the sample) is also limited.

The size of our sample, S, is made up of the sum of

$$\frac{S}{10} + \frac{S}{20} + \frac{S}{30} + \ldots + \frac{S}{10R};$$

that is, the frequencies of all the different words in a given sample, when added together, must constitute exactly the total number of running words in the sample. Expressed differently, these fractions must add up exactly to unity $\left(\frac{1}{1}\right)$. To illustrate this rather important point let us advance the hypothetical formula $F_r = \left(\frac{S}{2R}\right)$ which would yield the harmonic series

$$\frac{S}{2} + \frac{S}{4} + \frac{S}{6} + \frac{S}{8} + \ldots + \frac{S}{2R},$$

and which means that the most frequent word has ½ of all occurrences, the second ¼ and so on. Thus, for example, if we had a sample of 24 running words, the most frequent word would occur 12 times $\left(\frac{24}{2}\right)$, the second most frequent word 6 times $\left(\frac{24}{4}\right)$, and the 3rd and 4th words respectively 4 and 3 times. But the sum of 12 + 6 + 4 + 3 occurrences is 25, or one more occurrence than is permissible in our sample of 24 words. Therefore we may say that our hypothetical formula is valid for only 3 plus different words. It makes no difference how large the sample is; the four fractions $\frac{1}{2} + \frac{1}{4} + \frac{1}{6} + \frac{1}{8}$ will always add up to more than unity.

Obviously as we increase the size of the constant modifying R in the harmonic series, we thereby decrease the size of each successive fraction and increase the number of fractions in the series which will be necessary to add up to unity; that is, we increase the number of possible different words in a given sample. It might be instructive to see in tabular form how rapidly the number of different words (n), necessary to saturate the formula, will increase as one increases the size of the constant (K) modifying R (cf. Table III).

TABLE III

THE NUMBER (n) OF DIFFERENT WORDS (APPROXIMATE) NECESSARY TO SATISFY THE SERIES $K = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{n}$ FOR THE INTEGRAL VALUES FROM K = 5 TO K = 10 (OR 1/5R TO 1/10R)

Size of constant	Number of different words (approximate)
(1/KR)	(n)
1/5R	31
1/6R	227
1/7R	610
1/8R	1,670
1/9R	4,550
1/10R	12,370

If we may speak in terms of "saturation," we may say that our Formula I is "saturated" with approximately 12,370 different words. In this connection I call attention to the excellent analysis of Dr. Joos ('36, p. 200) where important aspects of our formulae are considered, and also to my answer ("Statistical Methods" '37, p. 63f) which treats certain aspects of Dr. Joos's analysis which to my mind are mistaken.

Now clearly in a sample of, say, only 24 running words, our Formula I will not manifest itself; the most frequent word, $\frac{24}{10}$, could occur at the most only 3 times, and the second, $\frac{24}{20}$, at the most only twice, while the remaining fractions could not repeat at all in so small a sample. That is, one word would occur 3 times, one word 2 times, and 19 words would occur only once. Our Formulae I, II, and III would not hold if only because the sample is too small. The exponent of F in Formula III would be much greater than the square; yet as we increased progressively to a certain point the size of the running sample, the exponent should, theoretically, come progressively nearer that of Formula III; and *mutatis mutandis* with Formula II. The smallness of the size of the sample, then, may affect the size of the exponents (discussed somewhat Zipf '35, p. 44 and '37 *passim*). Since we are interested in very large samples in this paper, we dismiss our consideration of very small and small samples.

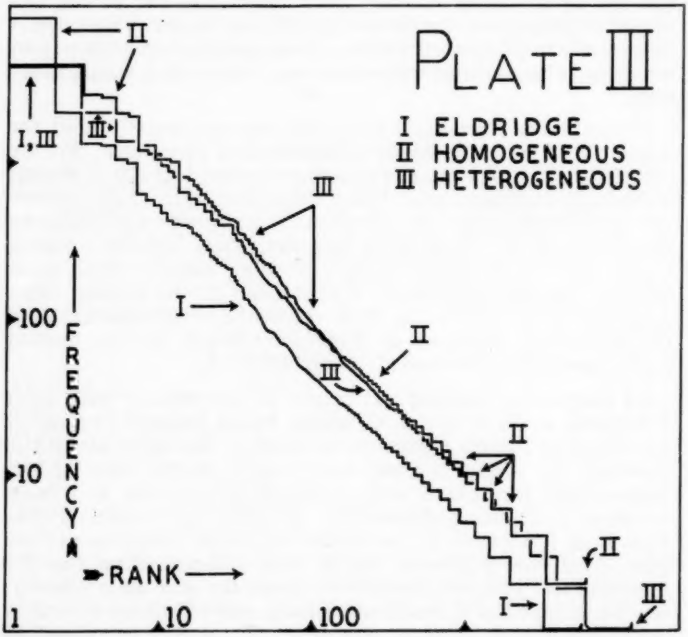
Nevertheless Formula I may also not *necessarily* hold for a sample of running words of indefinitely large size. For by restricting the size of R (to approximately 12,370 different words), it also says that these approximately 12,370 different words will have an opportunity to appear according to this formula in a sample of approximately 100,000 running words (and incidentally it says nothing more). That is, it imposes the sole restriction that R shall not be greater than approximately 12,370 (in this connection see the detailed mathematical analysis of Professor M. H. Stone, quoted Zipf, "Statistical Methods," '37, p. 63f.)

Of course, by shifting our origin in accordance with such a formula as $(R + a)F = C$, which would amount to moving the curve of Figure I parallel to itself to the right along the abscissa, we might include many more words than 12,370. Indeed this formula is very valuable empirically as I hope to show in future publications. But for the present I dismiss it as irrelevant to the matter at hand, since no matter how much we shift our origin thus (disregarding for the moment the obvious restriction upon the size of *a* above), our Formula III will continue to hold, since it is not stated in terms of R. And Dr. Thorndike doubts the general validity of Formula III.

We have now reviewed the statistical background of our general problem and shall now turn to a consideration of Dr. Thorndike's observations.

(1) The size of the exponent of Formula III *may* indeed vary between samples of different sizes as I have already remarked both here and elsewhere (Zipf '35, p. 43f.); hence I have protected myself against Dr. Thorndike's type of finding. On the other hand the exponent does not *have to* vary necessarily with increased size of samples because of anything inherent in our formulae, as I shall now show by considering two extreme cases of combining samples of equal sizes.

Eldridge's Newspaper count covered 43,989 running words and, as plotted according to Formula II (Zipf '35, Plate IV), gives a remarkably close fit. I present these data again in our accompanying Plate III; of the three curves drawn, it



is the one (marked I) nearest the origin. Let us now ask what would happen to our formulae if we added another sample of the same size, making thereby a combined aggregate of 87,978 words. Obviously the results will depend entirely upon the structure or frequency-distribution of the second sample.

For example, if the second sample were precisely the same as the Eldridge count in the sense that the same words occurred with the same rank and the same frequency, the combined samples would follow the formula log R + log 2F = log 2C. That is, we would simply have doubled the frequency of every ranked word in the Eldridge list; graphically we would have but moved our curve upward along the ordinate parallel to itself by an amount equal to the log of twice the frequency of any given ranked word minus the log of its frequency. Every bend in the Eldridge curve will be reflected in a corresponding bend in the curve above; the treads for frequencies of 1, 2, 3 etc. in the Eldridge data will appear at 2, 4, 6 etc. in the combined samples; and our Formula III would have to be changed accordingly by substituting $\left(\frac{F}{2}\right)^2$ for F^2. No word would occur once or have an odd frequency of occurrence. One of the curves (marked II) to the right on Plate III is this combined curve plotted.

We shall say, by definition, that two sets of data are completely *homogeneous* when upon combination they do not increase the number or change the order of the ranked words in either sample, but simply increase the frequency of each ranked word by a constant proportion. Thus two samples, A and B, of equal length, and each separately yielding precisely the same curve on double log grid, are completely homogeneous when upon combination the frequencies of the ranked words of either sample are simply doubled.

Two samples, A and B, of equal length, and each separately yielding precisely the same curve on double log grid, are completely *heterogeneous* when, upon combination, the number of different words is doubled and no ranked word in either sample changes its frequency. For sample, if A is the Eldridge material and B a frequency-count of 43,989 running words of Turkish or Hottentot or some other very exotic language in which the ranks and frequencies of ranks are precisely the same as those in Eldridge, with no two words in the two samples being phonetically identical, we should have an example, upon combination, of complete *heterogeneity*. The most frequent rank would be doubled and the effect of the combination would be to push the curve to the right parallel to itself along the abscissa by an amount equal to the log of $2R_n$ minus the log of R_n in which R_n is an arbitrary rank in either A or B. Formula III would still hold. This curve, labelled III, is also plotted on Plate III. Every bump or bend on the curve I has a precisely corresponding bump or bend to the right on curvè III, for all that we have done is to double the membership of each rank. The significant difference between II and III in respect to I is that II is above I, and III to the right of I.

An example of completely homogeneous language would be the speech of a person whose only verbalization was the constant repetition of, say, the Nicene Creed; verbalizations of this type are not unknown in psychotic material. Similarly homogeneous are the utterances of persons speaking a set piece in unison. For heterogeneity, consider the speech of an interpreter, or the editorial columns of different foreign-language newspapers published in this country.

We have seen from our two extreme cases that the simple formula, $2RF = 2C$, describes at least two entirely different types of material; the curves for II and III on Plate III will themselves give necessary information as to slopes and congruency as we now dismiss them.

And between the two extremes of complete homogeneity and complete heterogeneity as defined above, are degrees of "more or less" homogeneity, or partial homogeneity, whereby, upon combining of samples, ranks and frequencies only partially fall together. Tendencies towards convexity upwards or downwards in a given curve at certain of its portions would seem to indicate a "homogeneity" or "heterogeneity" respectively at those portions, whatever these terms may mean dynamically in their final analysis. The point we are now making, however, is that in combining samples of the same or different sizes one may at present make no predictions from our formulae as to the results without first having evaluated the samples to be combined in respect to their degree of homogeneity according to objective critera of the type above described. Since Dr. Thorndike does not make clear that he tested his samples as to homogeneity before combining them, we may dismiss his paper as having imposed no essential restrictions upon our formulae. I suspect that material cannot be tested for homogeneity or degrees of homogeneity, except by use of formulae of the type under consideration. There is a certain danger, then, of arguing in a vicious circle.

Of course, if we assume our formulae, we have valuable tests of homogeneity both in the speech-stream of an individual and in the speech-production of a social group. In the speech-continuum of an individual we would then see limits, bends, tendencies towards periodicity and the like; in the speech-production of a social group or social groups we would have some indication of what makes for likeness and what for dissimilarity, and thereby gain insight into what, if anything, is meant objectively by such statements as, say, "what cannot live together harmoniously does well to part company" or, say, "the female mental processes always have been, are, and forever will remain essentially incomprehensible to the male." But the value of any deduction from our formulae, or similar empirical formulae, will depend entirely upon the validity of such formulae as accurate descriptions of facts. Our immediate (but not sole) task then is to test our formulae in very diverse types of material. In this connection I mention that independently or in collaboration with others I have either completed or have practically completed analyses of the speech-production of children, American Indians, psychotics, and of samples of English at different periods of its development, and of samples of very different speech-groups in respect to their morpheme-distribution, the frequency-distribution of their loan-words, etc. These results will be made common property shortly, and the analyses continued in kind so that there will be ever less doubt about the applicability of our formulae one way or the other. Though theoretically we may never entirely escape the risk of the vicious circle mentioned at the end of the last paragraph, any charge thereof may be lodged, theoretically, equally well against the inductive-deductive methods of science in general; for in the last analysis we are but observing that the events of speech are not random in their occurrence and are investigating objectively the conditions which govern them, even as any other scientist.

In saying *peccavisti* to Dr. Thorndike in the matter of homogeneity of samples, I must also say *peccavi*, for I too have made a tacit assumption in the matter of homogeneity, without any regrets. It is inconceivable that Mr. Joyce, in writing his *Ulysses*, at no time lifted his pen from his paper; in any case one may view the continuum of *Ulysses* as a combination of parts, in respect to one person's writing. Similarly the Eldridge material, really an aggregate of newspaper clippings, may safely be viewed as a combination of samples of different persons' speech-production. And we did note in each case a close approximation to a linear equation similar for both no matter what our *a priori* assumptions in the matter may or may not have been. This leads us to the next point.

(2) In demonstrating our formulae, either here or previously, we have not begged the question statistically; hence our formulae are not "statistical artifacts" as that term is commonly understood. Furthermore, in view of the number of different tests to which the formulae have been submitted empirically, one cannot say that the correspondences are but matters of chance, as the word "chance" is commonly interpreted; indeed the probabilities are overwhelmingly against our finding such a close correspondence in respect to our formulae as appears in even the two arbitrarily selected samples of language represented by the Eldridge and Joyce materials respectively.

Furthermore I do not agree with Dr. Thorndike that those things added together lose much of their instructiveness in the combination. On the contrary, things added together may gain in instructiveness for reasons suggested above. Of course, the random search for coefficients of correlation, or the simple feeding of masses of data into calculating machines, may be considered an uneconomical procedure and an indication of a foolish unimaginative mind though not necessarily of a mistaken mind, for many a valuable thing has thus been stumbled upon, and many a problem would seem to admit of no other approach. Nevertheless research is likely to be more pleasant and productive if one has a hunch about where and how to look. To-day dynamic philology has rudiments of theory sufficiently well-founded to work with, though not to romp with on the playground of subjectivity. We know already that speech, when viewed as populations of acts, behaves according to law; we have good reason to suspect that the size of the population is an important factor in our observations; we are therefore warranted in studying various sizes of populations for what we may learn about the general effect of size. Theoretically, in a field evidently so complex as that of speech-production, we do well to approach our studies free from all bias except this: like things under like conditions remain alike. To define "things" operationally, and to determine "conditions" empirically constitute our essential task. Nevertheless in practice we already have reason to believe that there is a fundamental saturation-principle in operation in the continuum of speech, and we therefore do well to test empirically the concepts of over-saturation and under-saturation, attempting to correlate, say, differences in hypothetical degrees of saturation with other observable facts (to that end my own research mentioned above). It is to my mind not inconceivable, as I hope someday to show, that though the stream of speech activity may be viewed as a *continuum,* like a straight line or the circumference of a circle, it is *limited* in the manner of the circumference of a circle and unlike a straight line (we remember the limit of our Formulae I); and it is *bent,* unlike either a straight line or a circle, but generally like a spiral in the sense that a point revolves around an axis as it continuously and at a definite rate approaches it [we remember the ever-present marked trend in the direction of *abbreviation* of size, etc. (Zipf '35, *passim*)]. That is, the stream-of-speech winds up in an orderly fashion and never repeats. This is but an analogy; for we are not dealing with a single axis or a single point in two dimensions, but with populations of acts apparently organized throughout in many groups in many dimensions, and ordered to a very considerable extent in respect to relative (or comparative) frequency of occurrence. The analogy is but my own personal very general working hypothesis, here presented with neither the intention nor desire of enlisting a following. And yet the statement of some such analogy seems both legitimate and necessary in a paper where the combination of parts is under discussion. Small parts of a great spiral of wire could of course be fitted together "more or less" into either a straight piece of wire, or a number of circles, with either everything straight or everything circular, the deviations being conveniently charged off to "statistics." But in our case it is perhaps in the "more or less" that the vital factor lies. We cut a small section from the continuum of speech and indeed find a straight line as in our Plates I and III; but a huge section, or a huge aggregate of small sections would seem to tell a different story. Then, I believe, we should find a law of diminishing proportions of new increments in operation. This belief would have been belied had Dr. Thorndike found our Formula III in operation in samples totalling 4½ million running words. As it is his findings support it. Indeed were we to take ever larger samples of a given speech-continuum, plotting them always to our Formula II (e.g. Plate I), the left of our curve would probably rise proportionately with the increase logarithmically; but the bottom of the curve would turn down ever more. That is, with ever larger samples, the negative slope at the bottom would be ever steeper, its tangent approaching and finally becoming perpendicular to the abscissa. A curve of such a shape (almost typical of the language of little children, and otherwise frequent) is of the type Dr. Thorndike is running into; already the slope of his lowest 16 points, in terms of Formula II, is approximately –2 instead of –1 as I calculate them. And I believe that much of his discussion of Formula III would appear as I have suggested if it were in terms of Formula II. Furthermore, by superimposing rectangular coordinates on the double log grid on which is plotted our hypothetical enormous set of data according to Formula II, we might well find for it a close approximation to a parabola, or in any event a second degree parabola (in the statistical acceptation of the term) with a marked linearity for only a comparatively small portion at the top which need not disturb us now. Differentiating this hypothetical quasi parabola, we should have a formula telling us of the slope at any *R;* the reciprocal of this slope would be our "rate of variegation," i.e. the rate (ever in the direction of zero) at which variegation would be increased, or new words would be added, at that point. One might also plot the cumulative fractions of Formula I against increasing rank with substantially the same results, expressed, however, in terms of the size of the sample. I agree with Dr. Thorndike that ultimately no (or only negligibly few) new words would be added. But I also believe that in the continuous speech of an individual, and even perhaps in the combined speech of a closely-knit group, our "rate of variegation" would approach zero according to a formula of general applicability. Indeed what else may we suspect from even our data previously presented (and they are not unique)? In cases where one found marked deviations from a constant formula, one would doubtless find "things happening" in the structure or dynamics of speech to bring it into conformity. What else may one argue, in view of the fact that we find Formula III operative in genetically related but morphologically and formally different speech-continua (of comparable size) such as for American newspaper English, Plautus' Latin, and the words of the *Iliad* (to be reported separately)? When one considers what may "happen" to words phonetically, morphologically and semantically, one gets an idea of the enormous complexity of the total phenomenon of word-distribution; but from enormous complexity one may not argue randomness, or imponderability or immeasurability. It would be naive for us to-day to assume a spiral, or the threads of a screw, or something similar as an analogy; it would, however, be equally naive of us to-day, in view of the implications of our findings, to proceed oblivious to the possibility of this as an analogy. Other analogies would doubtless do equally well; a harmonic series would be *nulli secundum,* if we knew more about limits, or if we learned more about the dynamics of curve-smoothing in respect to speech, as we are indeed learning by studying English at various stages of its development, while remembering the "more or less" aspect of our criteria of homogeneity. True, a harmonic series is cold and impersonal. But in general terms we have but suggested, primarily in respect to language, that as we grow older we proportionately decrease the rate at which we extend the horizon of our activity—a proposition with which few would care to quarrel. The proposition in language is susceptible to empirical testing, though the task may be arduous because of the magnitudes involved; we remember, however, that by carefully weighing small pebbles at hand one may eventually learn how to weigh enormous masses at vast distances. The task is not the less worth undertaking because, for reasons to be somewhat clarified as we close, the answer may shed welcome light on many a philosophical quandary about the nature of life and death in general if only because speech cannot, and *must not,* be permanently abstracted from the rest of animate organization and behavior.

Since this paper is not intended to be a point-for-point rebuttal, but rather a clarifying and constructive answer to a very constructive paper, let us approach another point intimately connected with our three formulae: statements may be incomplete without being necessarily false. Dr. Thorndike has failed to show, I think, that any of the above three formulae are false. Let us ourselves show that they are all incomplete, as indeed they are, if only for one reason apart from the inadequacies just discussed. For if we turn our three formulae around, in the spirit of Jacobi's "man muss immer umkehren," we can by no means reconstruct the stream of speech in respect to a fundamental matter of word-distribution, even after taking into account the general limitations of all statistical formulation. Thus we do not know from our formulae whether words occurring, say, 20 times in a given sample, will occur one right after the other in immediate succession, or be distributed throughout the sample according to the principle of probability or some other principle. On this point our formulae are simply noncommittal. There may then be a formula of repetitions beyond our formulae of rank-frequency-number.

If one takes the words occurring, say, 20 times in the index of *Ulysses,* and computes for each different word, the number of page-intervals between recurrences (19 intervals for each word), disregarding line-references and classifying intervals only as one page or less, two pages, three pages etc., one notes, upon combining all like intervals for all words, that the intervals are distributed according to the approximate formula $IH = Constant$, in which I = the length of intervals in pages, and H (Häufigkeit) = the number of times the interval of each size occurs. In short, small intervals appear often, and large intervals rarely; and the size is apparently inversely proportionate to the frequency. If one takes 'a given ranked interval (e.g. the first, the third or the eighteenth interval) for all the words occurring 20 times, one finds substantially the same relationship. As one measures the intervals between repetitions of words occurring less often (say 10 times), one finds the same fundamental relationship as above, except that the constant is larger and the variability somewhat greater. These observations are made by inspection, to a first approximation, and subject to the corrections of additional studies now being made. It would seem that the size of the interval-unit selected may not be a matter of indifference. I here express my thanks to my former student, Mr. Alexander Murray Fowler, for painstakingly computing intervals for the words occurring 5, 10, 15, 20, and 24 times in *Ulysses;* I also call attention to the related work of Dr. B. F. Skinner ('36, p. 94).

The value of our new tentative formula is that our previous formulae might well be but corollaries to it. The tendency it apparently reflects, to my mind essentially the effect of "habit" or of the "cumulative force of chance" (Zipf '35, pp. 197-201), might be considered a "tendency towards perseveration," using in *perseveration* (following Skinner '36, p. 94), a psychological term descriptive of a frequently observed phenomenon.

The purpose of presenting our new type of analysis here is to intrude the thought that we are not necessarily committed in our future work to study frequency and variety only by operating with our present formulae. Dr. Thorndike (pp. 405-406) states: "I fear, therefore, that before we can use the relations between frequency of use and the number of words having each frequency safely in either linguistics or psychology, we will have to determine it in many instructive cases, increasing the sample in each case until the relation becomes stable." On the one hand that may be true for linguistics or psychology; in any event increasing the sample will be very instructive for science, even though we already have, I think, enough material to suggest that the frequency-number relation will not necessarily become stable (except as a regularly decreasing rate of variegation) upon increasing the size of the sample, if I understand Dr. Thorndike's words correctly, and that it (the rate of variegation) will not be the same for all types of language. On the other hand, we should do well to consider the possibility of discovering other, easier, and more instructive types of analysis, in the hope of superseding our present formulae with better ones. Indeed the genuine instructiveness of our present formulae seems in no small measure to be in the direction of suggesting more refined approaches and of formulating old and new problems more precisely.

(3) I believe (Dr. Thorndike remarks about my belief) that the equilibrating forces suggested by our formulae are wider and deeper than Dr. Thorndike does, and that they are acting to produce equilibrium for equilibrium's sake. In so far as we know anything of natural processes we see what is apparently equilibrium for equilibrium's sake everywhere. One could almost assume the same for speech *a fortiori*, though I have never operated with this *a fortiori* assumption expressed or implied.

After all, one may in no sense abstract the stream of speech or anything which is traditionally or actually a part of speech-behavior from the rest of our being; Dr. Thorndike's terms "speaker-writer relationship" or "hearer-reader relationship" have never been shown to represent entities. Indeed, if anyone is inclined to believe that his stream-of-speech is not connected with any given part of his anatomy, let him but jab a red-hot needle into that given part and observe a possible repercussion therefrom upon his speech-production; and *mutatis mutandis* with other portions of, or situations in, his experience, be it biological, psychological, or sociological (the terms, to my mind, in no sense discrete).

I am inclined to believe that the organization of living process may well be found to be essentially linear (with certain restrictions), in the sense that living process in itself proceeds according to ratios; and that the laws of living process may well be found to be isomorphic with those of speech. I am led to this belief by what seems to me the logical consequences of the apparent laws of speech themselves, glimpsing a certain unity of organic process apart from any cult on the unity of science. This belief, but a speculative hypothesis, may easily be mistaken and is certainly not here urged upon others; nevertheless it already takes on a remarkable robustness if one views the stream of speech, rightly I believe, as the resultant of a great many forces in operation to which we are reacting (’35, Chap. VI). The linearity of speech-habits is striking, even in latent speech and associated words (Skinner ’36, ’37); my unpublished studies of the morpheme-distribution of Gothic and Chinese indicate for great portions of the distribution the formula $\log F + R =$ a Constant (to be treated separately); the size of a phoneme doubtless depends to a great degree inversely upon the relative frequency of its occurrence (Zipf ’35, Chap. III). And yet the laws of speech are not exclusively linear. As predicted (ibid.), the Zwirners (’36, ’37) are finding that the distribution of speech-sounds (or speech-utterances) about a phonemic norm in significant cases and in respect to significant aspects of articulation (e.g., length of sound) follow the normal curve. I have observed that in the several languages examined the lengths of words (in phoneme-units) occurring once, twice, thrice, etc. are apparently distributed according to the normal curve with the mean ever more to the left as the frequency becomes greater (to be reported separately). Another type of curve, non-linear, then does occur. We must not overlook the fact, however, that in the actual utterance of speech, as, say, in the articulation of a phoneme, there is a problem of moving masses of tissue, that is, physical masses, in order to give expression to ideation; and that the Zwirners *may* be but noting the precision with which these masses are generally moved, and are but finding a formula frequently descriptive of the effect of man-moving-masses in trying to obtain an objective. But I do not see how this possible explanation will serve in the second apparently normal curve, the unit being the phoneme; for here differences in length are not “accidental” but definitely connected with differences in function-meaning (to be discussed separately).

Thanks to the investigations of a great many persons in the field of speech, and not least of all to those of Dr. Thorndike, we now have a fairly rich body of material with which to work. And all agree that the knot of speech should be untied with great care, precision, and objectivity. Time alone will tell whether we shall gain thereby very considerable insight into the way possibly similar knots are to be untied elsewhere, and nothing in this paper, or in Dr. Thorndike’s paper, I am sure, is to be construed to the contrary.

REFERENCES

Eldridge, R. C., '11—*Six Thousand Common English Words,* Buffalo.

Hanley, M. L., '37—*Word Index to James Joyce's Ulysses.* Madison, Wisconsin.

Joos, M., '36—Review of the Psycho-Biology of Language. *Language,* 12:196-210.

Skinner, B. F., '36—The Verbal Summator and a Method for the Study of Latent Speech. *Jour. Psychol.,* 2:71-107.

Skinner, B. F., '37—The Distribution of Associated Words. *Psychol. Rec.,* 1:69-76.

Thorndike, E. L., '37—On the Number of Words of any given Frequency of Use. *Psychol. Rec.,* 1:397-406.

Zipf, G. K., '35—*The Psycho-Biology of Language,* Boston.

Zipf, G. K., '37—Statistical Methods and Dynamic Philology. *Language,* 13:60-70.

Zipf, G. K., '37*—Observations of the Possible Effect of Mental Age upon the Frequency-Distribution of Words from the Viewpoint of Dynamic Philology. *Jour. Psychol.,* 4:239-244. *(Referred to by Dr. Thorndike.)

Zwirner, E., & Zwirner, K., '36—*Grundfragen der Phonometrie; Reihe A and B,* Berlin.

Zwirner, E., & Zwirner, K., '37—Phonometrischer Beitrag zur Frage der neuhochdeutschen Quantität. *Archiv f. vergleich. Phonetik,* 1:96-113.